Structural bioinformatics and molecular simulations

Looking at membrane proteins

Membrane proteins account for approximately 25% of all genes, and constitute approximately 50% of potential drug targets. The steady increase in the number of three-dimensional structures for membrane proteins means that the twin disciplines of structural bioinformatics and biomolecular simulations may be applied to this important class of molecules. Bioinformatics studies are starting to reveal, for example, sequence motifs that govern how transmembrane α-helices pack together. Simulations are revealing the dynamic behaviour of membrane proteins and the nature of their often transient interactions with the surrounding lipid molecules.

by Jonathan Cuthbertson and Mark S.P. Sansom
(Oxford, UK)

Membrane proteins are vital to many aspects of cellular function. However, we know relatively little concerning the structures of membrane proteins: only about 80 structures of membrane proteins have been determined at high resolution (see http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html for a summary), in contrast with a total of approximately 25 000 in the Protein Data Bank for non-membrane proteins. Bioinformatics and related techniques, such as molecular modelling and simulation, therefore play a central role in maximizing our understanding of membrane protein structure and function.

Structural bioinformatics and prediction

Although only a small number of structures are known, we can start to analyse common features of these structures. The majority of membrane proteins have a relatively simple architecture, being based upon a bundle of hydrophobic transmembrane (TM) α-helices. (The major exceptions to this are the bacterial outer membrane proteins, which are based upon a TM β-barrel architecture.) More recent work analysing the distribution of amino acid residues within membrane proteins[1] has added to the initial model of a simple hydrophobic TM helix. In particular, the importance of amphipathic aromatic residues (i.e. tryptophan and tyrosine) and of basic residues (lysine and arginine) has been recognized. These two classes of residue form 'bands' on the surface of membrane proteins, corresponding to the location of the lipid headgroups with which the side chains are believed to interact (see below).

The recognition of the importance of hydrophobic TM α-helices early on has resulted in a plethora of methods for prediction of TM helix topology based upon analysis of membrane protein sequences[2]. Earlier methods used simple hydrophobicity scales to search for potential membrane-spanning helices. More recent methods use sophisticated pattern matching methods (e.g. neural nets and hidden Markov models) to search for TM helices based on sequences of such elements in membrane proteins of known structure. These methods are quite successful (around 85%) at identifying the number and topology of TM helices within a membrane protein sequence. However, they vary in the efficiency with which they predict the exact beginning and end of TM helices (Figure 1A). There are now approximately 400 TM helices of known three-dimensional structure. This enables us to assess the relative efficiency of the various methods in predicting start and end residues (Cuthbertson *et al*, unpublished work).

As the predicted proteomes of fully sequenced organisms are now available (http://www.ebi.ac.uk/proteome/), genome-wide prediction of membrane proteins[3] has become relatively straightforward, facilitating comparative studies. In undertaking such studies, it is important to first filter sequences through a signal peptide predictor (e.g. SignalP[4]) in order to avoid false-positive predictions of N-terminal TM helices. By then combining the output of multiple TM

Key words: homology modelling, membrane protein, molecular dynamics simulation, molecular modelling, transmembrane α-helix.

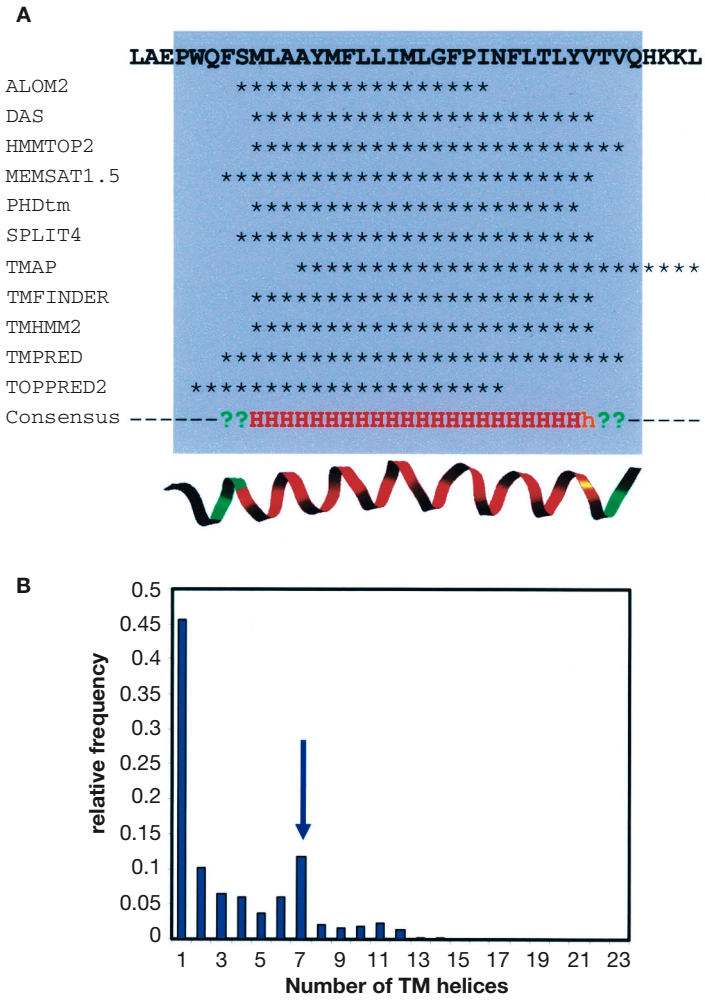

Figure 1. (**A**) Comparison of prediction methods for TM helix 1 from bovine rhodopsin. The sequence is given at the top. The blue box indicates the extent of the helix in the X-ray structure. For each method, the predicted TM helix residues are indicated by stars. The consensus prediction is given at the bottom, along with an image of TM helix 1 coloured as follows: red (H) implies that more than eight methods predict a TM residue; orange (h) implies more than five methods; and green (?) implies more than one method. (**B**) Predicted distribution of TM helices for the human genome. The peak at seven TM helices (arrow) corresponds to the G-protein-coupled receptors.

helix prediction methods, it is possible to provide consensus predictions for all the potential membrane proteins in a particular proteome. Consensus prediction for the >28 000 proteins identified to date reveals that roughly 23% of human proteins contain at least one TM helix. The most frequent class of membrane proteins is that with a single predicted TM helix (Figure 1B). There is a significant population of predicted membrane proteins with seven TM helices, which corresponds to the large family of G-protein-coupled receptors.

Current structural bioinformatics studies on membrane proteins are focusing on more complex sequence and structural motifs. Such motifs provide clues as to the packing together of TM helices within bundles, and to possible dynamic functional roles of TM helices. Two classes of sequence motif that have received particular attention: glycine-containing motifs (e.g. GXXXG) within TM helices and proline residues in the middle of TM helices. Senes et al.[5] have shown that the GXXXG motif occurs at a high frequency in TM helices, providing a means for mediating close interactions between adjacent TM helices. More recently, molecular modelling has been used to show how a homo-oligomeric bundle of TM helices, each containing a GXXXG motif, may form a transbilayer pore[6].

Proline-containing TM helices are found in a wide range of membrane proteins[7]. The presence of a proline residue usually distorts the TM helix to form a kink (Figure 2). Such kinks may act as 'molecular hinges', as suggested by a number of modelling and simulation studies. By providing a potential hinge within an otherwise relatively rigid TM element, proline kinks may play an important role in mechanisms of signalling and of transport across cell membranes.

Homology modelling

Homology modelling is playing an increasingly important role in studies of membrane proteins, especially when coupled with molecular simulation (see below). The reason for this is that as a result of difficulties in high level expression of eukaryotic membrane proteins, structures are more likely to be solved for bacterial homologues. A prominent example of this is the bacterial potassium channel KcsA, which provides a model for the core pore-forming domain of mammalian potassium channels. Such bacterial membrane protein structures may then be used as templates for modelling the structures of biomedically significant mammalian homologues.

The quality of a homology model depends upon the quality of the underlying sequence alignment (between the template and the novel target protein). Once an alignment has been optimized, a homology model can be readily generated using programs such as MODELLER (http://salilab.org/modeller/modeller.html) or SWISS-MODEL (http://www.expasy.org/swiss-mod/SWISS-MODEL.html). Of course, these methods have been developed and optimized with water-soluble proteins in mind, but they may be used for membrane proteins, especially if constraints derived from the results of TM

prediction programs and from experimental data are used to aid model evaluation and refinement.

As an example of the application of homology modelling to membrane proteins, we may consider the development of a model for the transmembrane domain of GluR0. GluR0 is a prokaryotic homologue of mammalian glutamate receptors (which form cation-selective ion channels in the central nervous system). When expressed in mammalian cells, GluR0 forms potassium-selective channels that are activated by glutamate. Sequence comparisons and TM predictions of GluR0 reveal that it has a TM domain homologous with that of KcsA. The channel is formed by a tetramer, the four subunits being arranged symmetrically around a central pore. Two equally plausible alignments of the sequences of the TM domains of KcsA and GluR0 may be constructed. These alignments differ in the exact position of the four-residue insertion in the GluR0 sequence in the vicinity of the P-helix that precedes the selectivity filter of the channel. Two closely related homology models of the GluR0 TM domain tetramer were subsequently constructed and compared using molecular dynamic simulations[8], allowing the more favourable model to be identified.

Molecular simulations

The crystal structure of a membrane protein represents an average (both spatial and temporal) structure, generally in a non-bilayer environment (e.g. in the presence of detergent molecules or bound to antibody fragments), and at a temperature of 100 K. In order to help us understand the function of membrane proteins in a lipid bilayer environment

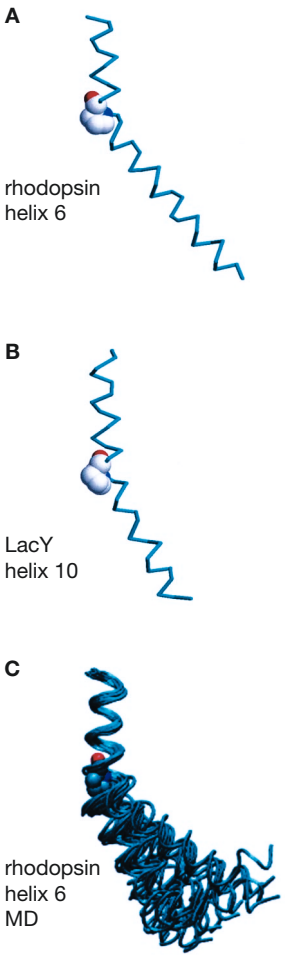

Figure 2. Proline-kinked TM helices. (**A**) Helix 6 from the X-ray structure (PDB code 1L9H) of bovine rhodopsin; (**B**) helix 10 from the X-ray structure (PDB code 1PV6) of lactose permease (LacY); and (**C**) superimposed structures from a multi-nanosecond simulation of helix 6 of rhodopsin showing potential flexibility about the proline hinge (D'Rozario and Sansom, unpublished work). (In each case the C-terminus of the helix is uppermost and the proline residue is shown in space-filling format.)

at a physiological temperature (approximately 300 K), we may use molecular dynamics (MD) simulations. Such simulations enable us to explore the conformational dynamics of a membrane protein on a ~10 ns timescale. The system simulated usually consists of a membrane protein embedded in a simple phospholipid bilayer [e.g. dimyristoylphosphatidylcholine (DMPC) or palmitoyloleoylphosphatidylcholine with water and ions on either side of the membrane. MD uses classical mechanics based on an empirical inter-atomic potential energy function to generate trajectories of atomic co-ordinates and velocities for successive time steps via numerical integration of the forces acting on all atoms in a system. In the past, MD simulations were impossible without access to supercomputing facilities. However, with the advent of (relatively) cheap Linux clusters, MD simulations of increasingly complex biomolecules are becoming widespread and provide a valuable complement to structural bioinformatics and experimental structural biology.

As an example of the use of MD simulations to understand the conformational dynamics of a membrane protein, we may consider the case of the N-terminal domain of the outer membrane protein OmpA from *Escherichia coli.* This eight-stranded β-barrel is one of the simplest structures of an outer membrane protein from a Gram-negative bacterium. MD simulations have enabled us to compare the behaviour of this protein in three different environments (Figure 3): in a lipid bilayer; in detergent micelles (as used in NMR structure determination); and in detergent-containing crystals. On the basis of 10 ns duration simulations, the protein was revealed to undergo dynamic fluctuations roughly 1.4 times larger in the micelle environment than in the bilayer, which were sufficient to reveal spontaneous openings of a central pore that is functionally closed in the crystal structure[9]. The fluctuations of OmpA in the crystal simulations were markedly less than in the bilayer environment (by a factor of approximately 1.3). These changes in flexibility are correlated with opening and closing of a continuous pore through the centre of the β-barrel, which explains the observed pore formation by this protein when reconstituted into lipid bilayers.

Lipid–protein interactions play a key role in folding and stability of membrane proteins. Crystal structures of membrane proteins usually

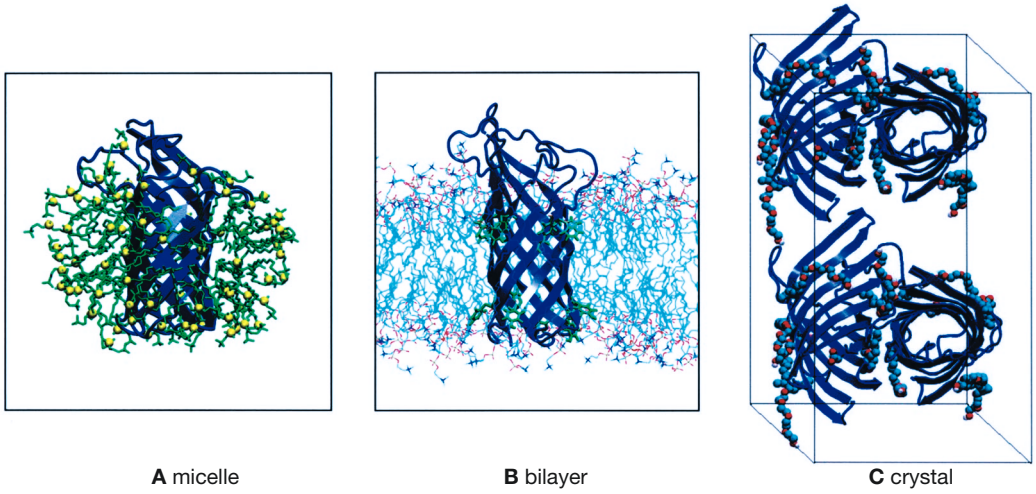

Figure 3. Images representing simulations of OmpA in three different environments. (**A**) A dodecylphosphocholine (green/yellow) micelle; (**B**) a dimyristoylphosphatidylcholine (cyan/red) bilayer; and (**C**) a crystal containing four OmpA molecules and 24 C8E4 (detergent; cyan/red spacefill) molecules per unit cell.

contain at most a single tightly bound lipid or detergent molecule, thus making structural bioinformatics studies of such interactions possible, but difficult (Deol and Sansom, unpublished work). MD simulations enable us to probe lipid–protein interactions more transient than those observed in crystals. For example, a comparison of simulations of OmpA and KcsA, both in phosphatidylcholine bilayers, emphasizes the role of aromatic belt residues in forming interactions with the polar headgroups of lipid molecules[10]. Recent simulations of the outer membrane protein OmpT (Baaden and Sansom, unpublished work) reveal a potential lipid-A-binding site on the surface of the β-barrel formed by a cluster of basic side chains. Thus simulations can provide valuable insights into the roles of lipid–protein interactions in membrane structure and stability.

Future directions

It is evident that the number of high-resolution structures of membrane proteins is steadily increasing. The impact of this on structural bioinformatics will be to provide statistically better sampling of structural motifs. This will enable us to improve prediction of membrane protein structures, in particular enabling us to move towards prediction of aspects of three-dimensional structure, such as the packing of TM α-helices within a lipid bilayer environment.

From a biomolecular simulation perspective, the availability of more structures will enable comparative analysis of the conformational dynamics of membrane proteins. This will be of particular interest in, for example, revealing common patterns of interactions of membrane proteins with the lipid molecules that surround them. Another important development is that simulations of more complex membrane protein systems will become possible, helping us to build a bridge between molecular and cellular level descriptions of, for example, membrane transport mechanisms. Such higher-level simulations will require development of more coarse-grained representations of membrane proteins and of lipids, enabling simulations to be performed on much longer time scales and much larger length scales than is currently feasible.

References

1 Ulmschneider, M.B. and Sansom, M.S.P. (2001) Biochim. Biophys. Acta **1512**, 1–14
2 Möller, S., Croning, D.R. and Apweiler, R. (2001) Bioinformatics **17**, 646–653
3 Wallin, E. and von Heijne, G. (1998) Protein Sci. **7**, 1029–1038
4 Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) Protein Eng. **10**, 1–6
5 Senes, A., Gerstein, M. and Engelman, D.M. (2000). J. Mol. Biol. **296**, 921–936
6 Kim, S., Chamberlain, A.K. and Bowie, J.U. (2004) Proc. Nat. Acad. Sci. U.S.A. **101**, 5988–5991
7 Cordes, F.S., Bright, J.N. and Sansom, M.S.P. (2002) J. Mol. Biol. **323**, 951–960
8 Arinaminpathy, Y., Biggin, P.C., Shrivastava, I.H. and Sansom, M.S.P. (2003) FEBS Lett. **553**, 321–327
9 Bond, P. and Sansom, M.S.P. (2003) J. Mol. Biol. **329**, 1035–1053
10 Domene, C., Bond, P.J., Deol, S.S. and Sansom, M.S.P. (2003) J. Am. Chem. Soc. **125**, 14966–14967

Jonathan Cuthbertson is a graduate student in the final year of his DPhil in Department of Biochemistry at the University of Oxford, where he is jointly supervised by Declan Doyle and Mark Sansom. His research is focused on bioinformatics approaches to membrane proteins, encompassing both prediction methods and molecular simulations.

email: cuthbert@biop.ox.ac.uk

Mark Sansom is a professor in the Department of Biochemistry at the University of Oxford, where he is head of the recently formed Structural Bioinformatics and Computational Biochemistry Unit. He received his degree from Oxford in 1983, and worked for 7 years in the University of Nottingham, before returning to Oxford. He has interests in many aspects of computational studies of membrane proteins (website: http://sansom.biop.ox.ac.uk) with special interests in ion channels, bacterial outer membrane proteins, and the development of high-end computing for simulation and modelling of complex membrane systems.

email: mark.sansom@biop.ox.ac.uk